APPLICATION OF

JONATHAN M. FRIEDMAN

FOR LETTERS PATENT OF THE UNITED STATES

A METHOD FOR *ab initio* DETERMINATION OF MACROMOLECULAR CRYSTALLOGRAPHIC PHASES AT MODERATE RESOLUTION BY A SYMMETRY-ENFORCED ORTHOGONAL MULTICENTER SPHERICAL HARMONIC-SPHERICAL BESSEL EXPANSION

James J. DeCarlo
Registration No. 36,120
Attorney for Applicant
STROOCK & STROOCK & LAVAN LLP
180 Maiden Lane
New York, New York 10038
(212) 806-5400

Our Docket No. 389004/039

A METHOD FOR *ab initio* DETERMINATION OF MACROMOLECULAR CRYSTALLOGRAPHIC PHASES AT MODERATE RESOLUTION BY A SYMMETRY-ENFORCED ORTHOGONAL MULTICENTER SPHERICAL HARMONIC-SPHERICAL BESSEL EXPANSION

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims priority of U. S. Provisional Appl. Ser. No. 60/219,863, filed July 20, 2000 under 35 U.S.C. §111(b).

FIELD OF THE INVENTION

The invention pertains to the field of using computational methods in predictive chemistry. More particularly, the invention utilizes techniques in crystallographic molecular replacement for drug design and *ab initio* molecular phasing. The techniques rely on a software program with associated algorithmic functions, to optimize the prediction of the crystallographic phases and structure for molecules of interest including proteins or other molecules have therapeutic value.

BACKGROUND OF THE INVENTION

The roles of medicinal chemist and crystallographer have not been altered in several decades. Their efforts to identify the structure of chemical compounds and therefrom deduce their chemotherapeutic effects, thereafter devising more potent or less toxic variations of them for medicinal use, has long been one involving the arduous task of attempting to crystallize and test one compound at a time to determine individual bio-activity and efficacy. This system is made even more costly and time consuming by the fact that over 10,000 compounds must be individually tested and evaluated for every compound that actually reaches market as a chemotherapeutic agent, World Pharmaceutical News, 01/09/96, (PJB Publications). These facts have driven many scientists and pharmaceutical houses to shift their research from traditional drug discovery (e.g. individual evaluation) towards the development of high throughput systems (HTP) or computational methods that will bring to bear increasingly powerful computer technology for the drug discovery process. To date none of these systems

have been proven to significantly shorten discovery and optimization time for the development of chemotherapeutic agents.

Accordingly, a need exists to optimize the prediction of bio-activity in chemical compounds such that the discovery and development of therapeutically valuable compounds is made more rapid and efficient.

SUMMARY OF INVENTION

Described here are details about, simplifications for, and enhancements to the accuracy of our recently described method [*Computers & Chemistry*, 23, 9-23 (1999)] for determining *ab initio* phases of macromolecular crystallographic structure factors at any experimental resolution limit. To apply this method, one first finds points in the unit cell that can serve as centers for large nonoverlapping spherical asymmetric units and chooses one such point, $\mathbf{x}_o$, as the origin of a set of spherical harmonic-spherical Bessel (SHSB) basis functions, $S^{\ell mn}(\mathbf{x}_o,r,\phi,\theta)$. The complex-valued Fourier space representation, $T^{\ell mn}(\mathbf{x}_o,hkl)$ of each real space basis function, $S^{\ell mn}(\mathbf{x}_o,r,\phi,\theta)$ for one asymmetric unit is combined, by complex summation with the crystallographic symmetry related Fourier space representations of the remaining asymmetric units, to create the Fourier space representation of a joint SHSB basis function $[F_{solo}^{\ell mn}(\mathbf{x}_o,hkl)]$ that can serve as a component basis function to describe the contents of an entire unit cell. The coefficient of each component function in the full-cell SHSB expansion is determined by a weighted linear least squares procedure. Given here is a more detailed explanation of this least squares procedure, a description about the general behavior of the coefficient refinement that enhances the speed of the calculation by about 2 orders of magnitude, a description of a "zonally restricted" packing function for selecting the origin for component basis functions, a method for extricating the refinement process from local minima, a statistical evalution of the refined *ab initio* phases that are produced for one specific test case at moderate resolution, and a presentation of typical electron density maps that are obtained for the medium resolution (2.7Å) phasing of tetragonal Staphylococcal nuclease.

DETAILED DESCRIPTION

In a previous paper, we outlined a method for the *ab initio* phasing of sparsely packed (macromolecular) crystals by transforming the problem of phasing into one of finding complex expansion coefficients for that linear combination of symmetry constrained orthogonal models, which is optimally consistent with the experimental diffraction pattern. We described a useful choice of such non-overlapping symmetry-expanded orthogonal functions for which the number of required coefficients scales well with resolution; that is, the number of independent parameters to be determined does not greatly exceed the number of experimentally determined diffraction data for any choice of experimental resolution range.

This advantage arises because our method does not presume an atomic model and thus does not require high resolution data for adequate experimental data to parameter ratios. Earlier *ab initio* methods may have suffered from assumptions of atomicity or of dense packing of atoms that are difficult to maintain at the low experimental resolution and with the sparse packing typical of macromolecular structures. A further advantage for choosing the SHSB basis functions is that the resulting expansion is relatively insensitive to reasonable choices of the origin. The initial disadvantage of the method was the amount of time required for the calculation. For example, our initial calcuation for the tetragonal form of Staphylococcal Nuclease required 9 wk on 16 nodes of a parallel processing IBM SP2 computer. We describe here some observations about the initial calculations have allowed us to reduce the computation time by between one and two orders of magnitude. For the Staphylococcal Nuclease test case, the time required for one cycle of the calculation was reduced from 9 wk to 2 d. This shorter calculation time has allowed us to optimize the accuracy of the procedure for this test case.

We wish, now, to elucidate upon methods by which one may obtain reliable convergence in the determination of a_{bnn}, the complex coefficients of the alternative expansion, from an experimental diffraction pattern. We wish also to describe our application of these methods to

determine *ab initio* phases for several proteins of known structure. Ultimately, here, we wish to provide a convincing demonstration of the utility of the electron density derived by these methods.

Overview of the Method:

Although the values of the coefficients of a SHSB expansion may vary with the choice of origin, the fidelity of the reconstructed image does not depend on the choice of origin, provided that the non-zero portion of the expanded 3-dimensional function lies *completely* within each of the chosen spherical zones of expansion (Fig. 1a). Thus, if one wishes to find a "symmetry-enforced" orthogonal expansion of the contents of a crystallographic unit cell in terms of SHSB basis functions, one may partition the unit cell into crystallographically symmetrically related spherical zones of expansion— one such zone for each asymmetric unit (Fig. 1b).*

If a SHSB expansion is chosen, it would be convenient to describe the largest possible portion of the unit cell as a linear combination of these SHSB basis functions. Bearing in mind that these SHSB functions are identically zero outside of the zones of expansion, the origin for each asymmetric unit may be placed at a point in the unit cell that is far away from all points related to itself by crystallographic symmetry (Hendrickson & Ward, 1976). The radius is then chosen to avoid overlap between adjacent spherical zones of expansion. Such overlap would cause degeneracy of the best fit solution and this degeneracy might hinder convergence to a unique solution.

Given an appropriate choice of radius and origin for the SHSB zones of expansion, then *at most* between 45% and 55% of the unit cell's contents may be represented by the expansion. Macromolecular crystals generally have a solvent content of greater than 45%, or a macromolecular content of *lower* than 55% (Matthews, 1968xxx). Furthermore, the intervening solvent regions

* Any similar complete set of orthogonal basis functions that avoids overlap between independent asymmetric units would suffice. However, if the basis set is chosen to be plane waves restricted to an *entire* asymmetric unit, *i.e.* the symmetry adaptation of a typical Fourier basis, then our method will break down because each plane wave basis function will be found to contribute only into a single reflection. This same feature of Fourier transforms gives rise to Heisenberg's uncertainty principle in quantum mechanics (Cohen-Tannoudji, *et al.*, 1980). The more extensive the region is that we wish to describe in direct space, the less extensive is the region of Fourier space from which the corresponding information is available (and *vice versa*).

can often be considered to be featureless (Wang, 197xxxx). Thus this choice of partioning between described and undescribed regions of the macromolecular unit cell may adequately account for a large portion of the macromolecular contribution to the x-ray diffraction pattern. The failure to account for all of the space in the unit cell dictates that a certain portion of the macromolecular electron density may lie outside of the zones of expansion and will thus fail to be accounted. (*i.e.* Some unaccountable electron density will inevitably fall into the null space of this SHSB basis.) However, an appropriate choice of SHSB origin is expected to minimize the amount of this undescribed density (Hendrickson & Ward, 1976).

Given known phases for a crystallographic diffraction pattern, a unique SHSB expansion is obtained that reproduces the expanded 3-dimensional image with high fidelity (Friedman, 1999). Without known phases, but with a known diffraction amplitudes, one may try to approach a self-consistent set of phases by successive approximations. Even if such an approach leads to convergence, one must anticipate that convergence may result in one of several trivially related isometric solutions. These related solutions can be converted into each other by some well known formulae that are listed below, and electron density calculated from each choice of solution can be analyzed for consistency with expectation.

Isometric Solutions:

We were initially concerned that macromolecular diffraction patterns might not represent the contents of a unique unit cell. Thus far, the only solutions that have arisen by our method are ones related to some of the expected alternate solutions.

Some alternative distributions of electron density, $\rho'(xyz)$, are expected to give rise to an experimental diffraction pattern that is identical to the diffraction produced by the actual crystal, except for differences in the values of the phase of each reflection. For instance, the photographic negative image of the unit cell gives rise to a diffraction pattern for which the calculated amplitude of each reflection is identical with the corresponding amplitude calculated for the true unit cell contents, but for which the phase of each reflection is different by 180 degrees. Likewise, the

amplitudes of reflections from the enantiomeric unit cell are identical with calculated amplitudes for the true unit cell, but with the phase of each reflection different by a sign.

A third class of alternate solutions for many space groups are those that are related by an arbitrary translation of the unit cell origin. Here, again, these equivalent alternate choices of origin lead to identical diffraction intensities, but the phase of each structure factor F(h,k,l) differs by 360(hx+ky+lz) degrees, where (x,y,z) is the translation vector, in fractional coordinates, that relates the two equivalent unit cell origins. Any such choice of origin is equally valid, but for the best comparison of the agreement between two independent solutions, translation to a common origin, enantiomer and photographic image (positive or negative) is required. Thus it is expected that any *ab initio* phasing method might converge to a unique solution that differs from the true (or expected) solution, but from which the true solution can be easily obtained.

One concern is that linear combinations of these valid solutions may themselves be alternative valid solutions. This is not a concern for linear combinations of enantiomeric solutions.

Diagram xxx. The imaginary components of the combined amplitudes cancel, but the real components are additive. Thus although the initial ratio of |F1| to |F2| I 1:2, the linear combination F(1)+F(1)*; F(2) + F(2)* of the enantiomorphs gives an approximate final ratio of 1:1.

Linear combination of the complex diffraction pattern arising from different enantiomers yields combined diffraction amplitudes that are inconsistent with the diffraction pattern of either enantiomer by itself; the relative amplitudes will vary markedly with the extent of the combination. Linear complex combinations of the diffraction of the positive and negative image of the unit cell, on the other hand, are expected to differ only in the overall scale of the calculated amplitudes. However, as will be discussed below, our choice of basis functions causes such linear combinations of the positive and negative photographic image unit cells to correspond to variation of the contrast between the molecular asymmetric unit and the solvent.

It is expected that convergence to the true solution is as likely as convergence to the enantiomorphic solution. However, in pairs of space groups with a chiral arrangement of general positions (*eg.* $P3_1$ & $P3_2$, $P4_1$ & $P4_3$, $P6_222$ & $P6_422$), it is expected that one enantiomorphic solution is dictated by the prior selection of one of the pair of enantiomorphic spacegroups. In space groups without a chiral arrangement of general positions, it is possible that individually derived $a_{\ell mn}$ coefficients of different $S_{solo}^{\ell mn}(hkl)$ component basis functions correlate optimally with different crystal enantiomorphs. Even if this is the case, appropriate combinations of the component $S_{solo}^{\ell mn}$ functions are expected to have higher correlation with the electron density than inappropriate ones. The same is expected to hold in Fourier space so that that F_{obs} will have higher correlation $r(|F_{accum}| <-> |F_{obs}|)$ with internally consistent linear combinations of basis functions, $F_{accum}(hkl)$, for one of the two enantiomorphs. Inconsistent linear combinations between terms from different enantiomorphs will give combined $F_{accum}(hkl)$ functions with lower overall correlation *versus* the observed diffraction data when compared with combinations from a unique enantiomorph. In the absence of symmetry-derived crystal chirality, convergence to either unique enantiomorph is equally likely,† but prior selection of origin $\mathbf{x}_o$ may predispose the refinement to converge to one of the two enantiomorphs.

The linear combination of the true solution with one related to its negative image results in an image with a different overall scale factor. Since the Fourier space structure factor with the phase of the negative image lies along the same line on the complex plane as the structure factor of the true solution, linear combination corresponds to an adjustment of the contrast between the macromolecule and the solvent. Provided that featureless regions (presumed to be the solvent regions) of electron density in the experimental unit cell correspond to regions that lie predominantly outside of the zones of expansion, then convergence to the direct image is expected for those solutions with the larger values of $r(|F_{accum}| <-> |F_{obs}|)$. Convergence to the negative image may be encountered in densely packed crystals, for which the local absence of macromolecular

† We note that none of the SHSB basis functions is chiral but that chirality arises from combinations of two or more SHSB functions both with odd valued $\ell \geq 1$ *and* odd valued $m \geq 1$ *and* from which the SHSB coefficient phase angles $\alpha_{\ell mn}$ differ from one another by an angle *other than* an exact integral multiple of π radians.

electron density is more of a rarity than the local presence of ordered density. It may also result from inappropriately selecting the origin of the zone of expansion to lie in the very middle of a solvent cavity.

The key assumption of our method is that the choice of origin does not significantly affect the quality of the reconstruction, provided that the object for which the shape is being approximated lies predominantly within these spherical ranges. In the first test case that we examined, the symmetry-expanded models can account for about 80-90% of the non-solvent density in the $P4_1$ (uniaxial) unit cell of Staphylococcal Nuclease. If acceptance, at each stage of successive approximation, depends on the degree of cross-correlation between the observed diffraction amplitudes, $F_{obs}(hkl)$, and the continually accumulated calculated structure factor, $F_{accum}(hkl)$, then (1) an observed final high degree of cross correlation between F_{accum} and F_{obs}, and (2) observed convergence to corresponding phase sets from independent starting points both would suggest that the *de facto* choice of arbitrary unit cell origin by our procedure is one for which overlap between the strongly morphological region of crystallographic electron density and the spherical zone of expansion is automatically optimized. This is particularly important for uniaxial space groups, for which one coordinate axis is completely arbitrary, and for other space groups with several equivalent choices of origins. Similarly, increased effectiveness at describing the strongly morphological regions of the electron density may predispose the refinement to converge to that enantiomeric unit cell, which has a monomer with average coordinates closer to $\mathbf{x}_o$, the arbitrarily selected origin of expansion. However, it is not ruled out that weak cross-correlation with one of the alternative isometric solutions may still contribute to the overall noise level.

Zonally Restricted Packing Functions to Pick an Origin for the Basis Functions:

Our method requires that one pick an origin for the zone of expansion to be close to the average coordinate of a macromolcular monomer in the crystal. An exact match is not required. For the space group P1, any point in the unit cell is equally vaild, but an arbitrary coordinate other than the coordinate (0,0,0) is chosen to avoid a centrosymmetric arranagement of the SHSB basis

set in the crystallographic unit cell. For space groups other than P1, the origin was originally chosen to be that point in the unit cell which is furthest away from all points that are related to itself by crystallographic symmetry. This corresponds to the global optimum point of the Hendrickson-Ward packing function. A quick check of 5 different readily available crystal structures suggested that this choice allowed one to obtain an origin within 5Å of the average coordinate of the protein monomer.

A further, more detailed analysis, made possible by an earlier systematic classification of the oligomeric states of proteins in the Protein Database (ref xxx), showed several deficiencies in this procedure. Shown in Fig. XXX is a histogram of distances between the absolute packing function optimum and the observed average coordinate of each of those xxxx monomeric proteins in the structural database that crystallized in space groups other than P1. The distances reported in this histogram are those to the nearest symmetry related monomer in either the true or the enantiomeric unit cell, with consderation of all possible choices of unit cell origin. Clearly, distances greater than 20Å are expected to be insufficiently close for expansion zone radii on the order of 20Å to 40Å. To try to improve the selection of the origin, we considered local optima other than the absolute optima (Fig. xxx). This leads to some improvement, but still leaves a large percentage of crystal forms for which the closest of the top 20 peaks in the packing function still lies more than 12Å away from the average coordinate of the closest monomer.

Inspection of some of the poorer matches, led us to realize that the global optimum of the packing functions for some of these poor matches corresponds to a noteworthy position in the unit cell, but one that was in the very middle of a solvent channel rather close to the middle of a protein region. Further comparison of the average fractional coordinate vectors of monomeric proteins in macromolecular crystal forms belonging to the same Laue group suggested that unit cells in each Laue group contain certain "sweet spots." That is, the unit cell contains several points in fractional coordinates about which values for the average coordinate of the crystalline macromolecular monomers are clustered. Optima in zones about each of these points must considered seriously for a successful *ab initio* estimation of the average coordinate, even if the value of the packing function

is somewhat below the global optimum in these zones. Thus it appears that our difficulties arose from an often observed clustering of local optima near the absolute optimum of the packing function. The values of the packing function among these clusters of local optima near the global optimum are often sufficiently great that they can swamp out local optima in the other zones.

Thus a two stage search is conducted. In the first stage the values of the packing function are examined coarsely, only at each of the "sweet spots." In the second stage a finer search is conducted in independent regions near the top 20 (30%xxx) of the "sweet spots". Thus by imposing zonal restrictions, we mean that we are looking only for the local absolute maximum in each of the independent regions. The solutions found by this algorithm are distributed more evenly between the independent zones within the unit cell and one obtains the histogram of distances in Fig. xx. Each such 2-stage search takes an average of about 6s of real time using 16 parallel nodes on an IBM-SP2 computer. By using the zonal restrictions, then, one can get one point in the list of the top 20 to be within 5Å of the average coordinate of a monomer over 95% of the time. In practice, one may carry out the initial stages of SHSB coefficient refinement (*vide infra*) and select that origin which yields the largest low order coefficients as an appropriate choice of origin.

To summarize the results to this point, it is possible to describe a single ("monomeric") asymmetric object in space by a 3-dimensional spherical harmonic-spherical Bessel (SHSB) expansion:

$$(1)\ \rho_{monomer}(\mathbf{x}) = \Sigma_{\ell mn}\, a_{\ell mn}\, S_{monomer}^{\ell mn}(\mathbf{x}_o;\, r,\phi,\theta) = \Sigma_{\ell mn}\, |a_{\ell mn}|\, S_{monomer}^{\ell mn}(\mathbf{x}_o;\, r,\phi,\theta)\, e^{i\alpha_{\ell mn}}$$
$$= \Sigma_{\ell mn}\, |a_{\ell mn}|\, S_{mono}^{\ell mn}(\mathbf{x}_o,\, \alpha_{\ell mn};\, r,\phi,\theta)\ ,$$

where $\mathbf{x}_o$ is the selected origin vector. Once the proper origin is selected, the crystallographic unit cell is filled with nonoverlapping monomeric basis functions, each rotated and translated by crystal symmetry. This symmetry expansion of the monomeric basis functions yields $S^{\ell mn}_{solo}(x,y,z)$:

$$(2)\ S_{solo}^{\ell mn}(\mathbf{x}_o,\, \alpha_{\ell mn};\, r,\phi,\theta) = \Sigma_{sym}\, S_{mono}^{\ell mn}(\Re_{sym}^{\ x}\, \mathbf{x}_o + t_{sym},\, \alpha_{\ell mn};\, r,\, \Re_{sym}^{\ \phi}\, \phi,\, \Re_{sym}^{\ \theta}\, \theta)$$

the joint, full-unit-cell basis function. The effect of complex multiplication by $e^{i\alpha_{\ell mn}}$ is a rotation of the initial $S_{monomer}^{\ell mn}$ basis function by the angle ($\alpha_{\ell mn}$ /*m.*) prior to symmetry expansion. The task at hand, then, is to estimate the complex coefficients $a_{\ell mn}$ to obtain an estimate of

(3) $\rho_{unit\ cell}(xyz) = \Sigma_{\ell mn} |a_{\ell mn}| S_{solo}^{\ell mn}(\mathbf{x}_o, \alpha_{\ell mn}; r,\phi,\theta) = \Sigma_{sym} \rho_{mono}(\Re_{sym}{}^{x}\, \mathbf{x} + \mathbf{t}_{sym})$,

where $\Re_{sym}$ and $\mathbf{t}_{sym}$ correspond to operators that effect a unique crystallographic symmetry rotation and translation respectively.

We note that the $a_{\ell mn}$ coefficients in the above summations are complex numbers (*i.e.* $a_{\ell mn} = |a_{\ell mn}|e^{i\alpha_{\ell mn}}$), when $m \neq 0$. Since the Fourier transform is a linear transformation and since the basis functions have a finite range, the Fourier transform of this summation is the summation of the Fourier transforms of each of the components.

(3) $F_{unit\ cell}(hkl) = \Sigma_{\ell mn} |a_{\ell mn}| F_{solo}^{\ell mn}(\mathbf{x}_o, \alpha_{\ell mn}; r,\phi,\theta) =$

$= \Sigma_{\ell mn} \Sigma_{sym} |a_{\ell mn}| T^{\ell mn}(\mathbf{x}_o, \alpha_{\ell mn}; \Re^{T}_{sym}{}^{x}\, \mathbf{h})$

$= \Sigma_{\ell mn} \Sigma_{sym} |a_{\ell mn}| T^{\ell mn}(\alpha_{\ell mn}; \Re^{T}_{sym}{}^{x}\, \mathbf{h})\, e^{2\pi i(\mathbf{h}\cdot \Re_{sym}{}^{x}\, \mathbf{x}_o + \mathbf{h}\cdot \mathbf{t}_{sym})}$

Analytical expressions for the Fourier transforms of each of the component basis functions are known (Friedman, 1998; Crowther, 19xx; Dodson, 19xx), and thus one may construct a Fourier space combined basis function that represents a unit cell's worth of orthogonal basis functions. The numerical values of the SHSB basis functions were calculated by a robust recursion formula (ref) for which the m index varied the most slowly. This recursion is particularly convenient for this application because it permitted all $a_{\ell mn}$ coefficients with restricted phase values ($m = 0$) to be calculated before $a_{\ell mn}$ coefficients with less restricted phase value.

Estimation of SHSB Coefficients and Refinement of the Orthogonal Model:

The Fourier space full unit cell basis function, $F_{solo}^{\ell mn}(\alpha_{\ell mn}; hkl)$ (Fig. 2), corresponds to the phased, Fourier space representation of a unit cell that has been filled with non-overlapping SHSB basis functions, $S_{mono}^{\ell mn}(\mathbf{x}_o, \alpha_{\ell mn}; r,\phi,\theta)$, that are related by crystallographic rotational and translation symmetry. The choice of this class of basis function combined with the required absence of overlap between adjacent component real space SHSB basis functions, $S_{mono}^{\ell mn}$ leads to orthonormality of the $S_{solo}^{\ell mn}$:

(4) $\int_{unit\ cell} dV\, S^{*}_{solo}{}^{\ell mn}\, S_{solo}^{\ell',m',n'} = \begin{cases} N_{sym}, & \text{if } \ell=\ell', m=m', n=n' \\ 0, & \text{otherwise} \end{cases}$

That each corresponding Fourier space component function, $F_{solo}^{\ell mn}(hkl)$, is also orthonormal in the same sense follows from Parseval's theorem, which equates integrals of functions in real space to the integrals of their Fourier space functional representations. The scale factor that we want, corresponding to the scale of the experimental unit cell to a union of non-overlapping component functions, would be a summation over direct space of the point by point product between $S_{solo}^{\ell mn}$ (the union of direct space basis functions $S_{monomer}^{\ell mn}$) and the unknown crystallographic electron density. This is equivalent, within a sign, to the value of direct space convolution product at the single translation point $\mathbf{t}_0 = (0,0,0)$. It therefore follows, from the convolution theorem, that the amplitude of the desired $a_{\ell mn}$ coefficient is equal to the inverse Fourier transform of the point by point Fourier space product, but only at the position $\mathbf{x} = (0,0,0)$. To obtain this value of the direct space convolution product at the direct space position, $\mathbf{x} = (0,0,0)$, the Fourier kernel becomes equal to one and thus direct summation of the point by point product in Fourier space equals that in direct space. Unfortunately, an exact determination of $a_{\ell mn}$ requires prior knowledge of the phases of the Fourier space structure factors for the experimental electron density that is being expanded, because complex values must be used in the point by point Fourier space product. Thus, starting from diffraction amplitudes, the complex values of the coefficients $a_{\ell mn}$ may at best only be obtained by successive approximation.

Refinement of amplitudes $|a_{\ell mn}|$:

Our initial scheme to refine the orthogonal SHSB series model, in the absence of input phase information, was to use the current best estimates of the Fourier space phases and amplitudes at each stage in the calculation of subsequent coefficients. The idea was to use a refinement scheme that started with the determination of all SHSB expansion coefficients for which the value of the index m was 0. For these functions, the phase of $a_{\ell mn}$ is limited to be 0° or 180° by the physical requirement for non-imaginary values of the real space electron density (Fig. 3).

On the very first cycle and to a first approximation, we presume the *totipotency* of the symmetry expanded real space function S_{solo}^{001}. That is, we assume that S_{solo}^{001}, suitably weighted

and with an adequately chosen origin, $\mathbf{x}_o$, can by itself (solo) account approximately for all of the electron density that gives rise to the experimental diffraction. (For earlier work with similar assumptions compare Podjarny *et al.* 199x.) If the assumption of totipotency holds approximately, then we can start accumulating a set of estimated structure factors based on this:

$$(5)\ F^{0}_{accum}(hkl) = a'_{001} F_{solo}^{001}(hkl)\,.$$

To obtain an initial estimate of the coefficient a_{001}, we use the expression:

$$(6)\ a_{001} = \Sigma_{hkl}\, F^{*}_{solo}{}^{001}(hkl)\, F_{obs}(hkl) / (\Sigma_{hkl} F^{*}_{solo}{}^{001}(hkl) F_{solo}^{001}(hkl))\,,$$

which follows from the orthonormality of the F_{solo} functions and is equivalent to a least squares scale factor.‡ The normalization term in Eq. (6), $\{1 / \Sigma_{hkl}[F^{*}_{solo}{}^{\ell mn}(\alpha_{\ell mn};hkl)\ F_{solo}^{\ell mn}(\alpha_{\ell mn};hkl)]\}$, should remain constant, but is calculated explicitly at each index to avoid possible numerical errors. In practice, we have found it necessary to weight these initial estimates of the coefficient values by one minus the probability that the correlation between F_{obs} and $F_{solo}^{\ell mn}$ is random. Use of this weighted $a_{\ell mn}$ coefficient allows one to calculate the initial estimate estimate of the complex Fourier structure factors:

$$(7)\ a'_{001} = w(r_{Fobs\text{-}Fsolo})\ a_{001}$$

$$(8)\ w(r_{Fobs\text{-}Fsolo}) = 1 - \mathrm{erfc}\left[\frac{1}{2} \left| \ln\left(\frac{1 + r_{Fobs\text{-}Fsolo}}{1 - r_{Fobs\text{-}Fsolo}} \right) \right| \frac{\sqrt{N-3}}{\sqrt{2}} \right]$$

‡ Essentially, $F_{solo,\ell mn}(hkl,\alpha_{\ell mn})$ is the Fourier space representation of a SHSB joint basis function with a coefficient of unit modulus and an arbitrary phase. The question we ask is, "What is the proportionality factor between this basis function and F_{obs}, presuming that the phase of the SHSB coefficient ($a_{\ell mn}$) is $\alpha_{\ell mn}$?" It is presumed that the proportionality is all real and thus the imaginary part is a measure of the goodness of fit. In terms of linear least squares (Strang, 1976), the real part is the projection onto the space of possible outcomes and the imaginary part represents the distance (and direction) from this presumed model space.

On subsequent cycles (*eg.* cycle ν), we calculate a reduced structure factor, $F_{reduced}(hkl)$, to use in place of the unphased $F_{obs}(hkl)$ for comparison with $F_{solo}^{\ell mn}(\alpha_{\ell mn};hkl)$. Again we presume totipotency of $F_{solo}^{\ell mn}(\alpha_{\ell mn};hkl)$ in accounting for the *remaining* undescribed portion of the diffraction pattern ($F_{reduced}$) and scale each independent coefficient, in turn, by the following least squares relationship:

$$(9)\ F^{\nu}_{reduced}(hkl) = (|F_{obs}(hkl)| - |F^{\nu}_{accum}(hkl)|)\, e^{i\phi^{\nu}_{accum}}$$

$$(10)\ a_{lmn} = Re\{ \Sigma_{hkl} F^{*}_{solo}{}^{\ell mn}(hkl)\, F^{\nu}_{reduced}(hkl) / [\Sigma_{hkl} F^{*}_{solo}{}^{\ell mn}(hkl) F_{solo}{}^{\ell mn}(hkl)] \}$$

$$(11)\ a'_{lmn} = w(r_{Freduced\text{-}Fsolo})\, a_{lmn}$$

$$(12)\ F^{\nu+1}_{accum}(hkl) = F^{\nu}_{accum}(hkl) + a'_{lmn} F_{solo}{}^{lmn}(hkl)$$

Phases ($\alpha_{\ell mn}$) of the Expansion Coefficients $a_{\ell mn}$, the m=0 terms:

We always make use of prior approximations to the electron density by using calculated phases from each previous cycle as the best estimate for phases associated with complex Fourier space values. The values determined in the previous section only address the scale factors between $F_{reduced}$ and F_{solo}, for a single presumed value of $\alpha_{\ell mn}$, and thus only the amplitudes of the expansion coefficients $a_{\ell mn}$. When the value of the index *m* equals zero, $a_{\ell mn}$ is limited to values along the positive or negative real axis by the restriction that the unit cell contain completely real electron density. Physical intuition would dictate that, with a proper choice of expansion zone radius, choice of the expansion zone origin near to the monomeric center of mass (or average coordinate) should cause the value of the coefficient a_{001} to be large and positive. However, in our application, diffraction patterns F_{solo}^{001} corresponding to $\alpha_{001} = 0°$ and $\alpha_{001} = 180°$ are both stored

for further refinement. Our initial refinement scheme entailed saving accumulated diffraction patterns (F_{accum}) corresponding to as many combinations of the choices of $\alpha_{\ell mn}$, as was allowed by allotted computer memory. (Storage space for up to 16 independent F_{accum} functions was routinely available.) Once memory became exhausted, only those accumulated solutions F_{accum} with the top cross-correlation between $|F_{obs}|$ and $|F_{accum}|$ were retained. By refining the $m = 0$ terms first, in effect, we are first determining phases for a model that is presumed to be rotationally averaged about an arbitrary "z" axis, (which is arbitrarily chosen to coincide with the c-axis of the crystal for the initially calculated monomer).

Phases ($\alpha_{\ell mn}$) of the Expansion Coefficients $a_{\ell mn}$, the m≠0 terms (The Slow Calculation)

Comparison of the complex cross-correlation values is also carried over to those $a_{\ell mn}$ coefficients for which the values are not limited to be real. In this case, $F_{solo}^{\ell mn}(\mathbf{x}_o,\alpha_{\ell mn};hkl)$ in eqs. 6 & 8 again (xxx) is that diffraction pattern arising from a unit cell filled by crystallographic symmetry expansion of the direct space basis function $S_{mono}^{\ell mn}(\mathbf{x}_o,\alpha_{\ell mn};r,\phi,\theta)$. The argument $\alpha_{\ell mn}$ indicates that this full-unit-cell basis function is calculated by premultiplying the initial monomeric direct space basis function by $e^{i\alpha_{\ell mn}}$ prior to symmetry expansion and the argument $\mathbf{x}_o$ indicates the chosen origin of the expansion zone for this initial monomeric basis function. To select a value for $\alpha_{\ell mn}$, we initially calculated plots of $r_{solo\text{-}red}$ [*i.e.* the complex correlation coefficient between $F_{solo}^{\ell mn}(\mathbf{x}_o,\alpha_{\ell mn};hkl)$ and $F_{reduced}(hkl)$] versus the presumed value of $\alpha_{\ell mn}$. The unweighted modulus of the coefficient $a_{\ell mn} = A_{\ell mn}e^{i\alpha_{\ell mn}}$ is chosen to be the scale factor at one of the angular optima in the r *vs.* α plot. The computer program was initially set to consider weighted $F_{solo}^{\ell mn}(\mathbf{x}_o,\alpha_{\ell mn};hkl)$ functions for up to 16 of these optima with respect to $\alpha_{\ell mn}$. In this initial, slower calculation, we presumed, in turn, 72 values of $\alpha_{\ell mn}$, at 5 degree intervals, from 0 to 355 degrees inclusively, when m≠0. Because storage space was limited, two separate cycles were run. On the first cycle, $F_{solo}^{\ell mn}(\alpha_{\ell mn})$ was calculated for all 72 values of $\alpha_{\ell mn}$, and the r vs. α plot was calculated. Those with the best cross-correlation to $F_{reduced}$ were found and noted, but not stored. On the second cycle, these top 16 optima were stored and tried again with each of the 16 stored values of

F_{accum}(hkl). The maximum number of storage locations for F_{accum} (hkl) functions was a compile time parameter that could be changed arbitrarily. In the original version, we tested two different choices for this parameter and found that some significant solutions were discarded if only 8 of the F_{accum} (hkl) functions were stored at each cycle. The source code allowed distribution of the computation evenly among an arbitrary number of parallel processors for (1) the 1152 (= 72 x 16) test summations on Cycle 1, *i.e.* the initial plot of $r_{solo\text{-}red}$ *vs.* α_{tmn}, (2) for the 256 test summations on Cycle 2, and (3) for the initial least squares scale factor. Below we note some observations that now allow us to forego most of these comparisons.

The ultimately chosen value of α_{tmn} is that value which leads to the highest absolute value of complex correlation† between the basis vector F^{tmn}_{solo}(hkl) and the remnant "data" vector ($F_{reduced}$(hkl), the RHS vector). At each stage F_{accum}(hkl) is updated (Eq. 12) to include all prior knowledge from previous cycles. Also, cycle by cycle rescaling of F_{accum} to F_{obs} prevents the value of the the scale factor between these two Fourier space functions from wandering.

The a_{tmn} values determined as described above are only approximate, because the best estimate of the phases of the accumulated calculated structure factors (ϕ^{v}_{accum} in Eq. 9) at each cycle is also approximate. We wished to determine empirically whether such estimates of a_{tmn} could be refined by successive approximation to ϕ_{accum}. As described above, several F_{accum}(hkl) solutions were stored at each cycle for each combination between F_{accum}(hkl) from a prior cycle and $F_{solo}(\mathbf{x}_o, \alpha_{tmn}; hkl)$ with presumed values of α_{tmn} that gave rise to optimal cross-correlation. The intent of such a multisolution method was to circumvent the coarseness in the choice of α_{tmn} and to circumvent possible problems arising from accidentally high correlation between F_{solo} and isometric distributions of "remnant" electron density .

---

† This complex correlation is a correlation function between a paired list of complex numbers for which all product terms ($f_0 f_1$), in the normal definition of the correlation coefficient are replaced by the complex product ($f_0^* f_1$). In terms of the complex arguments (phase angles) ϕ_0 and ϕ_1:

$$r_{01} = \frac{n\Sigma[f_0 f_1 \cos(\phi_0 - \phi_1)] - \{\Sigma(f_0 \cos\phi_0)\}\{\Sigma(f_1 \cos\phi_1)\} - \{\Sigma(f_0 \sin\phi_0)\}\{\Sigma(f_1 \sin\phi_1)\}}{[\, n\Sigma(f_0^2) - \{\Sigma(f_0 \cos\phi_0)\}^2 - \{\Sigma(f_0 \sin\phi_0)\}^2 \,]^{1/2} \; [\, n\Sigma(f_1^2) - \{\Sigma(f_1 \cos\phi_1)\}^2 - \{\Sigma(f_1 \sin\phi_1)\}^2 \,]^{1/2}} .$$

Although the position of the basis function origin in the reconstructed, calculated unit cell is fixed, such "accidentally" high correlation between a single basis function [$F_{solo}^{\ell mn}(hkl,\alpha_{\ell mn})$] and poorly phased diffraction data may result from an inappropriate comparison with electron density in a unit cell for which the arbitrary origin, enantiomer, or photographic image differs. For proteins that crystallize in uniaxial space groups, such as Statphylococcal Nuclease, even for the right enantiomer and photographic image, accidental correlation may be found with electron density in a unit cell related by an arbitrary z-translation. Comparison of correlation coefficients between the observed structure factor amplitudes F_{obs} and a precombination $F_{solo}^{\ell mn}(hkl,\alpha_{\ell mn})$ with $F_{accum}(hkl)$ should allow fixing to a common origin. However, on preliminary cycles where ϕ_{accum} is poorly defined, the degree of inaccuracy in the current estimates of F_{accum} can still lead to inconsistency in the choice of origin.

Thus, the $a_{\ell mn}$ coefficients were improved recursively. The combined estimate of $a_{\ell mn}$ appears to become more well determined as the current overall estimated $F_{accum}(hkl)$ becomes better defined.

In this fashion, successive approximation was achieved but at a high cost in terms of CPU hours.

To avoid having the approximate nature of the ϕ_{accum} cause the optimization of $a_{\ell mn}$ to stray too far from the true solution, constant retracing (*i.e.* correction of previously determined values of $a_{\ell mn}$) was undertaken. Thus, in the initial slow calculation, before proceding to the next higher value of the m index (m_{new}), corrective approximation to $a_{\ell mn}$ was restarted from the index $m = 0$, and carried out over $a_{\ell mn}$ with all intervening values of m.

Observations from the slow calculation:

(1) The variation of correlation coefficients with presumed $a_{\ell mn}$ value is, in general, unimodally sinusoidal for basis functions with nonzero values of the m index. Typical plots of $r(F_{obs}\{- F_{solo}\} <->F_{solo})$ *vs.* $\alpha_{\ell mn}$ are shown in Fig. XXX and are overlaid with plots of the imaginary residual of

$A_{\ell mn}$‡ *vs.* $\alpha_{\ell mn}$. [to figure caption: To conserve disk space, the program is set to plot out only one of every five of the presumed phase angles that are actually considered for acceptance by the calculation.](Fix XXX). The scale factor is only approximately unimodal and is generally out of phase with the correlation coefficient sinusoid. Thus, rather than calculating scale factors and correlation coefficients for 72 independent presumed values of $\alpha_{\ell mn}$, it is only necessary to calculate initially those for 2 presumed values of $\alpha_{\ell mn}$, 0° and 90°. From these two values and an arc tangent function, we can find the $\alpha_{\ell mn}$ value at optimal correlation. This reduces considerably the amount of calculation power that is necessary; alone this improvement reduced the time from 9 weeks to less than 1 week.

(2) Convergence of the $a_{\ell mn}$ coefficients to > 95% stability is generally achieved after about 4 to 6 recursive cycles of refinement. Initially, we restarted from $m = 0$ before the initial calculation of coefficients for the next higher value of the m index (m_{new}), to avoid wandering. We find instead that one needs only restart the calculation from $m = m_{new}$-4 or $m = m_{new}$-5. We suggest that, for higher accuracy, the entire process should be restarted several times (at least twice) from $m = 0$; however, from analysis of the updated changes in coefficient values at lower m index (See *eg.* table XX), we find that we were initially overly conservative in the extent of reoptimization of coefficients for the lower order indices.

(3) The calculation may be skipped for those basis function for which the weighted coefficient is smaller than a set cutoff value. A convenient cutoff value is 10^{-7} times the value of the coefficient with the greatest absolute value of the coefficient a on a given cycle.

With the above improvements, the time required for fitting the 2.7Å Staphylococcal Nuclease data or the calculation was reduced from 9 wk on 16 nodes to 2 d on 4 nodes. This reduction in the time for the calculation of phases allowed us to vary several other parameters of the refinement to see whether obvious improvements could be obtained. At present, the reduction in the required number of comparisons, due to the sinusoidal dependence, leaves the initial parallelization scheme

‡ See the earlier footnote with this symbol.

inefficient if more than 4 nodes are used. Additional improvements in the parallelization are expected to improve the speed of the calculation even further. For problems with more moderately sized proteins and higher symmetry, the time for 1 cycle of refinement is still 1 to several weeks.

Electron Density Calculation:

The result of the SHSB expansion calculation is a set of reconstructed Fourier coefficients ($F_{accum}(hkl)$) that are continuously updated (accumulated) throughout the expansion procedure. These may be treated as a set of calculated structure factor amplitudes and phases in some of the generally used types of weighted difference Fourier maps. We initially tried to use σ_A wieghted $2F_o$-F_c style electron density maps (R.Reed xxxx), and were surprised to find that the optimal choice of σ_A resulted in maps for which the suggested weighting provided a $2F_c$-F_o map, rather than a $2F_o$-F_c style map. As expected, this leads to positive electron density for the region of the protein, within the confines of the spherical zone of expansion, and negative electron density in the regions outside of the expansion zone. These external regions are undecribed by the calculated model. The map which optimally matched the known test structure was a $2F_o$-F_c map using Sim weights (ref to Sim xxx).

One can rationalize this observation by noting that Sim's original derivation presumed that the sole source of error between F_{calc} and F_{obs} derives from missing atoms, *i.e.* electron density that has not been included in the present model. The derivation of the σ_A weighting scheme expanded upon Sim weighting by also accounting for positional error in the atoms that already have been included in the model.

Extent of the Spherical Harmonic Expansion Indices:

Different upper limits for indices ℓ, m, and n have been suggested by different authors for the description of centrosymmetric diffraction data. In the present application of the spherical

harmonic basis, we must achieve a compromise between maximal descriptive content and a minimal ratio of statistical parameters to number of experimental data. Several different choices of index limits were assessed for the case of phasing the $P4_1$ form of Staphalococcal Nuclease at 2.8Å (xxxx unique calculated diffraction amplitudes). These choices included:

(1) A full complement of ℓ and n indices but an artificially low cutoff in the index m to avoid underdetermination (xxxx data, xxxx SHSB amplitudes, xxxxx SHSB signs, xxxx SHSB phases).

(2) The full Crowther / Navazza cutoff for 2.8Å diffraction data (xxxx data, xxxx SHSB amplitudes, xxxx SHSB signs, xxxx SHSB phases.) It may be argued that the SHSB coefficient phases contain less information than the SHSB amplitudes because of their more restricted range of values. This trial choice of cutoff was chosen to demonstrate the effect of completely ignoring the low data to parameter ratio.

(3) The full Crowther / Navazza cutoff for $2.8*(2)^{1/3}$ Å diffraction data. This effectively reduces the resolution of the calculated diffraction pattern to that of a diffraction pattern that fills half of the Fourier space volume of the true experimental diffraction data. This allows the Fourier space values $|F_{cal}|(hkl)$ and $\phi_{cal}(hkl)$ to be determined by an equal number of experimental observations $|F_{obs}(hkl)|$.

Recursive Improvement of Initial Estimates of $a_{\ell mn}$:

Recursive improvement is accomplished by finding complex valued corrections to the initial coefficents by fitting $F_{solo,\ell mn}$'s to the complex difference, $(F_{obs}-F_{accum})$. Two different methods were examined for recursive improvement of the $a_{\ell mn}$ coefficients. In the first of these, initial estimates were determined for all coefficients before any recursive improvement was started. The second method involved recursive improvement of all indices up to index m-1, before any new coefficients of index m were determined. (Only the first cycle, at index *m=0*, lacked prior recursive improvement.)

After all coefficients with a given m index have been estimated, it is likely that the resulting F_{accum} is a better estimate of F_{expt} than the prior, less complete summations. Complex valued corrections are necessary due to the contributions arising from accidental correlation to alternative solutions in preliminary estimates of $a_{\ell mn}$.

The Computational Algorithm:

A flow chart of the algorithm is outlined in Fig. xxx. Several calculation modes have been incorporated into the program for convenience. Parallelization is crucial only to those calculation modes that determine cryatallographic phases from experimental amplitudes (modes 1 and 2):

Mode 1 f_{obs} -> f_{est}, maximum |r| is considered to be the optimum

Mode 2 f_{obs} -> f_{est}, maximum r is considered to be the optimum

Mode 3 f_{calc} -> $a_{\ell mn,calc}$ (known phases for f_{calc})

Mode 4 $a_{\ell mn,calc}$ -> f_{calc} (known phases for $a_{\ell mn,calc}$).

Empirical comparison of modes 1 and 2 reveals that mode 1 converges to solutions with higher combined overall correlation and chooses solutions that are more often consistent with minimal values for the imaginary residual in $A_{\ell mn}$. Recursive improvement is only required if complex phases are not known for either f_{calc} or $a_{\ell mn}$ coefficents. Thus no recursion or probabilistic comparison of correlation coefficients is required for modes 3 and 4.

(1) $a_{\ell mn} = \int_{r<a_{rad}} \rho(r,\phi,\theta)\, j_\ell(k_{\ell n} r)\, Y^{*m\ell}(\phi,\theta)\, r^2 \sin\theta \; dr\, d\phi\, d\theta$

The function $S_{\ell mn}(r,\phi,\theta) = j_\ell(k_{\ell n} r)\, Y^{*m\ell}(\phi,\theta)$

(2) $a_{\ell mn}(0,0,0) = N_{\ell m} \times$

$(-1)^\ell 4\pi k_{\ell n} (2a_{rad})^{1/2} \Sigma_h \; |F_h| e^{i(\psi_h - \pi\ell/2 - m\phi_h)} P^m_\ell(\cos\theta_h)\; j_\ell(2\pi R_h a_{rad})/(4\pi^2 R_h^2 - k_{\ell n}^2)$,

where $N_{\ell m}$ is a normalization term equal to sqrt{ [(2 ℓ + 1) (ℓ - m)!] / [4π (ℓ + m)!] }. In this

(3) $a_{\ell mn}(t_x,t_y,t_z) = N_{\ell m} \times$

$(-1)^{\ell} 4\pi k_{\ell n} (2a_{rad})^{1/2} \Sigma_h |F_h| e^{i(\psi_h - \pi\ell/2 - m\phi_h)} P^m_{\ell}(\cos\theta_h) j_{\ell}(2\pi R_h a_{rad})/(4\pi^2 R_h^2 - k_{\ell n}^2) e^{-2\pi i(Ht_x + Kt_y + Lt_z)}$

(4) $\rho(r_s,\phi,\theta,t_x,t_y,t_z) = \Sigma_{\ell m} c_{\ell m}(r_s,t_x,t_y,t_z) Y_{\ell m}(\phi,\theta),$

and the corresponding required coefficients are given by:

(5) $c_{\ell m}(r_s,t_x,t_y,t_z) = N_{\ell m} \times$

$(-1)^{\ell} 4\pi \Sigma_h |F_h| e^{i(\psi_h - \pi\ell/2 - m\phi_h)} P^m_{\ell}(\cos\theta_h) j_{\ell}(2\pi R_h r_s) e^{-2\pi i(Ht_x + Kt_y + Lt_z)}$

(6) $\rho(x,y,z) = \Sigma_{\ell mn} a_{\ell mn} S_{\ell mn}(r,\phi,\theta,t_x,t_y,t_z) = \Sigma_h F(h) e^{-2\pi i \mathbf{h}\cdot\mathbf{x}}$

(7) $F(h) = N_{\ell m} \times$

$(-1)^{\ell} 4\pi (2a_{rad})^{1/2} e^{2\pi i \mathbf{h}\cdot\mathbf{t}} \Sigma_{\ell mn} a_{\ell mn}(\mathbf{t}) e^{i(m\phi_h + \pi\ell/2)} k_{\ell n} P^m_{\ell}(\cos\theta_h) j_{\ell}(2\pi a_{rad} R_h) /(4\pi^2 R_h^2 - k_{\ell n}^2).$

‡

(9) $a_{\ell mn}(t_x,t_y,t_z) = N_{\ell m} \times$

$(-1)^{\ell+1} 4\pi (2a_{rad})^{1/2} \Sigma_h |F_h| e^{i(\psi_h - \pi\ell/2 - m\phi_h)} P^m_{\ell}(\cos\theta_h) (a_{rad}/2) j_{\ell+1}(k_{\ell n} a_{rad}) e^{-2\pi i(Ht_x + Kt_y + Lt_z)}$

‡ The appropriate integral for equations (9) & (10) is now equivalent to 5.54.2, p. 634 in Gradshteyn & Ryzhik (1980).

The original parallel algorithm for FAIZER used a single processor (node) for each combination of Fsolo and Faccum. If it were necessary to combine Fsolo's, each calculated with 72 different presumed values of the SHSB alpha angle, with 16 different stored lists of Faccum, then the 72 x 16 calculations could be split relatively efficiently between nodes. However, once it was found that only two choices of presumed alpha angles for the SHSB-coefficient for Fsolo were necessary for each calculation of a coefficient value, then the original parallelization scheme was found to be markedly inefficient. That is, combination of <u>two</u> choices of Fsolo (each having a value for the presumed alpha phase angle set at either 0 or 90 degrees) with <u>two</u> choices of Faccum, would have allowed at most <u>four</u> processors to be used efficiently for the calculation of scale factors and complex correlation coefficient values between Fsolo and Faccum-Fobs. Therefore, to speed the calculation further, parallelization was accomplished by splitting long summations efficiently between several nodes for the calculation of values of the {Faccum-Fobs,Phi.accum} <-> {Fsolo,Phi.solo} scale factor and for the calculation of the corresponding correlation coefficient. The program was modified to determine the most efficient splitting of each branch of the calculation between variable numbers of nodes, based on the number of nodes available and on the required number of branches of the calculation. For example, for Fsolos and Faccums each containing a list of 10,000 diffraction data, if 4 processors are available for a single calculation of a scale factor, the newly parallelized calculation will sum about 2,500 numbers on each processor and then combine the 4 partial sums afterwards, cutting run time for the calculation approximately by a factor of 4. The difficulty in achieving such parallelization is in maintaining that each partial summation within a branch of the calculation is combined with proper, corresponding branch members. Such proper communication was achieved with intra-communicator subroutines available from the MPI-Library. Further difficulty may arise if time required for internode communication begins to be similar to the time required for the calculation.

Chemical Representation:

Atoms in Molecules:

The ultimate representation to achieve.

Parameters:

a) x,y,z + uncertainty
b) thus 4 - 6 parameters for each atom

Limitations:

No overlap of adjacent, non-interacting atoms.

Advantage:

Direct interpretation in terms of chemical principles.

Plane Wave Representation

Linear Combination of Orthnormal Basis Functions:

Linear coefficients (F_{hkl}) available through cyrstallographic experiments.

Parameters:

One complex coefficient (2 parameters) for each plane wave.

Limitations:

For diffraction from a crystal, equivalent origin points in the unit cell must lie at the same position (phase) with respect to the cosine wave cycle.

Advantage:

1) They are directly related to experimental measurement.
2) Their geometry allows a complete description of the unit cell contents.

SHSB Expansion:

Fidelity of the SHSB representation of a 3-D object:

Insensitive to SHSB origin.

Choice of SHSB origin/radius:

a) to fill Maximum amount of space in a unit cell with non-overlapping, crystal symmetry-related SHSB functions.
b) each SHSB basis restricted to represent the molecular fragment for a single asymmetric unit of the crystal.

Intermediate Expansion Coefficients:

a_{lmn} from statistically-weighted least squares.

Data to Parameter Ratio:

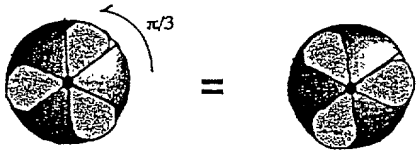
of a_{lmn} expansion coefficients = # of measured F_{obs}, at nearly every resolution range.
thus, #data / #parameters ≥ 1.00.

What about the phase angle (α_{lmn}) of the a_{lmn} coefficients?:

=> In general a_{lmn} is a complex number:

$$a_{lmn} = |a_{lmn}| \, e^{\,i\alpha_{lmn}}$$

=> Physically, the α_{lmn} correspond to a rotation of the starting basis functions by the angle α_{lmn}/m about the polar axis .

$e^{\,i\pi}$ x 

=

Slice through S_{331}
viewing down z-axis

=> However, since electron density is all real, the phase angles for coefficients, a_{l0n}, of the axially symmetric functions are limited to 0 or 180 degrees.

=> A further consequence of all real electron is that $a_{l,-m,n} = a^*_{lmn}$

Complex Cross Correlation vs. Presumed α_{111} for S_{111}

Cross-Correlation (r-complex) ⟶

"@" = r-complex > 0 ; "a" = r-complex < 0

analysis a(lmn) i n d e x: 1 1 1

	α	bin	lsq scale	r-complex	
P	0	12	-1.31676E+00	-1.40677E-01i	. . . . 1.= . . . s. a .
R	25	12	-1.49744E+00	-1.48911E-01I	. . . . 1= . . . . a
E	50	12	-1.36205E+00	-1.41854E-01I	. . . . .=1 . . . .sa .
S	75	12	-9.37860E-01	-1.18642E-01I	. . . . .= 1 .s . a. . .
U	100	10	-3.52441E-01	-7.40138E-02I	. . s . . a .=2 .1 . . . .
M	125	4	2.57782E-01	6.19706E-02I	. $. . @. 2 .= . 1 . . . .
E	150	8	8.11273E-01	1.11167E-01I	. . . . 2=$ 1. . @ . . .
D	175	12	1.25064E+00	1.37321E-01I	. . . . 2=1 . . . $.@ .
	200	12	1.48450E+00	1.48440E-01I	. . . . 2= . . . . . @
A	225	12	1.41519E+00	1.44525E-01I	. . . . 2= . . . . $@ .
L	250	12	1.04049E+00	1.24685E-01i	. . . . 1 2= . $.@ . .
P	275	12	4.74587E-01	8.54586E-02i	. . .$ .1 2.=@ . . . . .
H	300	4	-1.37341E-01	-4.45173E-02i	s .1 a . . .= . .2 . . .
A	325	8	-7.06905E-01	-1.03761E-01i	. . . . 1 s.= . a. . . .
	350	12	-1.17557E+00	-1.33338E-01i	. . . . 1 .= . . s. a. .

finding minima between which to sum

minimum # 1 occurs at position 24

minimum # 2 occurs at position 60

total # of minima found 2

top 256 of 2 correlation coefficients for lmn by itself

1 in angle register: 6 correlation ==> -.148911

2 in angle register: 42 correlation ==> .148911

# of selected pks: 2

TABLE I

Complex Cross Correlation vs. Presumed α_{111} for S_{111}

Cross-Correlation (r-complex) ⟶

"@" = r-complex > 0 ; "a" = r-complex < 0

analysis a(lmn) i n d e x: 1 1 1

	α	bin	lsq scale	r-complex
P	0	12	-1.31676E+00	-1.40677E-01i
R	25	12	-1.49744E+00	-1.48911E-01I
E	50	12	-1.36205E+00	-1.41854E-01I
S	75	12	-9.37860E-01	-1.18642E-01I
U	100	10	-3.52441E-01	-7.40138E-02I
M	125	4	2.57782E-01	6.19706E-02I
E	150	8	8.11273E-01	1.11167E-01I
D	175	12	1.25064E+00	1.37321E-01I
	200	12	1.48450E+00	1.48440E-01I
A	225	12	1.41519E+00	1.44525E-01I
L	250	12	1.04049E+00	1.24685E-01i
P	275	12	4.74587E-01	8.54586E-02i
H	300	4	-1.37341E-01	-4.45173E-02i
A	325	8	-7.06905E-01	-1.03761E-01i
	350	12	-1.17557E+00	-1.33338E-01i

finding minima between which to sum

minimum # 1 occurs at position 24

minimum # 2 occurs at position 60

total # of minima found 2

top 256 of 2 correlation coefficients for lmn by itself

1 in angle register: 6 correlation ==> -.148911

2 in angle register: 42 correlation ==> .148911

# of selected pks: 2

TABLE II

F_{accum}:

Rather than storing A_{lmn}, we store F_{accum}, "recalculated structure factors" that include phases.

This is accomplished by accumulaating the contribution from each SHSB basis function (i.e. from each lmn index) to F_{accum} at each step.

Electron Density Maps:

Standard Sim weighted 2Fo-Fc style maps may be calculated (where Fc is taken to be F_{accum}).

Degree of Convergence:

Compare the correlation coefficient between Fobs and Fcalc due to the orhogonal model (i.e. F_{accum}).

Some "Final" correlation coefficients

			Data	Model
Staph. Nuclease	$P4_1$	r = 0.95	Fcal	2.7Å
A DNA duplex	P321	4 = 0.85	2.2Å	2.7Å
A Recombinase/DNA	$P6_222$	r = 0.73	3.1/4.5	3.9Å

Sometimes alternative, non-equivalent origins are possible for the basis functions. For Staphylococcal nuclease, refinement, based on this alternative choice of origin led to a new set of Fcalc values, which upon translation to a common origin, had a complex cross-correlation of 0.81 with the set of Fcalc values from the original choice of origin.

Realizations:

Expansion of the spherical portion of a unit cell into SHSB expansions can be calculated by the convolution theorem. (Translation function)

$a_{.mn}(x,y,z)$, EACH GRID POINT HAS ITS
OWN EXPANSION IN lmn.
(Slow, but once)

Calculation of Empirical Energy Functions is a convolution (overlap integral).

Potential Function Component * ligand:

charge(x,y,z)
vdwA(x,y,z)
vdwB(x,y,z)

(Fast)

Structure Based Drug Design by Searching Through a Drug Database

1. The search problem is simplified to a 6-dimensional search of ligand positions and orientations.

2. A semi-exhaustive 6-dimensional search for the most stable protein-ligand configuration is made feasible by some tricks wit Fourier transforms and other orthogonal functional expansions.

3. Versions of these tricks have been used by crystallographers to find the orientation and position of known molecular structures in different packing configurations in new crystal forms.

Alternative Solutions

Non-Equivalent Origins:

For Staphylococcal nuclease, refinement, based on an alternative choice of origin, led to a new set of Fcalc values, which, upon translation to a common origin, had a complex cross-correlation of 0.81 with the set of Fcalc values from the original choice of origin.

Negative Photographic Image, Enantiomeric Unit Cell:

Staphylococcal nuclease ($P4_1$):

NO enantiomorphic soln.
YES negative image.

A DNA Duplex (P321):

? enantiomorphic soln.
NO negative image.

Either of these alternative solutions can be interconverted by addition of a consent to or negation of the calculated phase.

$$\rho(xyz) = \Sigma_{sym}\, S_{lmn}(\mathfrak{R}_{sym}, \mathbf{t}_{sym}, \mathbf{x}_o, r, \phi, \theta)$$

$$a_{001} = \Sigma_{hkl}\, F^{*}_{solo}{}^{001}(hkl)\, F_{obs}(hkl) / (\Sigma_{hkl} F^{*}_{solo}{}^{001}(hkl) F_{solo}{}^{001}(hkl))$$

$$a'_{001} = w(r_{Fobs\text{-}Fsolo})\, a_{001}$$

$$w(r_{Fobs\text{-}Fsolo}) = 1 - \mathrm{erfc}\left[\frac{1}{2} \left| \ln\left(\frac{1 + r_{Fobs\text{-}Fsolo}}{1 - r_{Fobs\text{-}Fsolo}} \right) \right| \frac{\sqrt{N-3}}{\sqrt{2}} \right]$$

$$F^{0}_{accum}(hkl) = a'_{001} F_{solo}{}^{001}(hkl)$$

$$a_{lmn} = \Sigma_{hkl}\, F^{*}_{solo}{}^{lmn}(hkl)\, [\,|F_{obs}(hkl)| - |F^{v}_{accum}(hkl)|\,] e^{i\phi^{v}_{accum}} / [\Sigma_{hkl} F^{*}_{solo}{}^{001}(hkl) F_{solo}{}^{001}(hkl)]$$

$$F^{v+1}_{accum}(hkl) = F^{v}_{accum}(hkl) + a'_{lmn} F_{solo}{}^{lmn}(hkl)$$

$$F^{v}_{reduced}(hkl) = (\,|F_{obs}(hkl)| - |F^{v}_{accum}(hkl)|\,)\, e^{i\phi^{v}_{accum}}$$

$$a_{lmn} = \mathrm{Re}\{\, \Sigma_{hkl}\, F^{*}_{solo}{}^{lmn}(hkl)\, F^{v}_{reduced}(hkl) / [\Sigma_{hkl} F^{*}_{solo}{}^{lmn}(hkl) F_{solo}{}^{lmn}(hkl)] \,\}$$

$$a'_{lmn} = w(r_{Freduced\text{-}Fsolo})\, a_{lmn}$$

SUMMARY OF THE METHOD

Everything is done on a grid. (Allows (FFT).

Find possible translation sites.

Expand the potential functions for each protein in terms of S_{lmn}. (A couple of hours).

Store the expansions of the spatial distribution of (charge/van der Waals) parameters for all drugs in a database. (A few days).

Fast searches for each drug using phased Crowther rotation search at each possible translation point. (Fraction of a second per site per drug).

The arbitrary choice of origin that is apparent from the application of spherical harmonic-Bessel expansions toward a six-dimensional search, and the high fideli6y for interconversion between the spherical harmonic-Bessel and Fourier representations suggest a method for describing the contents of a sparsely packed, non-centrosymmetric crystalline array in terms of multiple, non-overlapping, symmetry-enforced expansion zones. If all of the non-null electron density in a crystalline unit cell is contained within the limits of several non-overlapping spherical expansion zones placed into this crystalline cell, one may use the interconversion process to estimate the complex valued spherical harmonic-Bessel expansion coefficients from an incomplete Fourier description (diffraction amplitudes).

Each spherical harmonic-Bessel basis function of the representation can be used to generate an aggregate orthogonal basis function over a large portion of the entire unit cell. One applies crystal symmetry to rotate and translate an initial single-center spherical harmonic-Bessel basis function from within a single spherical expansion zone into several non-overlapping, crystal symmetry-related spherical expansion zones. One may multiply the initial basis function by a complex coefficient of unit amplitude and arbitrary complex phase prior to symmetry expansion. Conversion of the full unit cell aggregate spherical harmonic basis into the Fourier-basis results in a partial structure factor for index *lmn*. (In practice we calculate the same 'aggregate basis function' partial Fourier structure factor by first converting the initial single-sphere basis function tot he Fourier representation and then applying the symmetry.) For each choice of arbitrary spherical harmonic coefficient phase angle, the scale factor between this 'aggregate-basis function' partial Fourier structure factor and an experimental diffraction pattern gives an estimate of the amplitude of the true spherical harmonic-Bessel coefficient. The correlation coefficient between this first 'aggregate-basis function' partial Fourier structure factor and the experimental, incomplete Fourier representation (diffraction amplitudes) gives an indication of the goodness of fit. Differences in this correlation coefficient may be used to select an optimal complex valued spherical harmonic-Bessel coefficient from among several initially arbitrary choices of complex phase angles for the coefficient of the spherical harmonic-Bessel basis function. Thus, the amplitude of each spherical harmonic-Bessel coefficient can be chosen as the least squares scale factor between the aggregate basis function and the diffraction pattern; the complex phase of each spherical harmonic-Bessel coefficient can be chosen to be that which

optimizes the correlation coefficient between the Fourier representation of the basis function and the diffraction pattern. The orthogonality of the aggregate spherical harmonic-Bessel basis functions results in a lack of correlation between the coefficients calculated for the different component basis functions (i.e. for those with different values of the indices *l*, *m* and *n*). Thus, if all of the density in a crystal lies within expansion zones, one obtains a unique expansion. As this condition breaks down, there is expected to be a gradual accumulation of error in the diffraction pattern reconstructed from the spherical harmonic-Bessel basis. (The error arising from electron density outside of the expansion zones is exacerbated if the number of coefficients used in the spherical harmonic-Bessel expansion exceeds the number of available Fourier amplitudes.)

Because of the arbitrary nature of the origin for the expansion zones, the expansion zone can be chosen to be that which allows the maximum volume of the unit cell to be contained within non-overlapping expansion zones after symmetry expansion of the initial basis function. Up to about 55% of the unit cell's contents can be accounted for in this manner, a percentage commensurate wit the non-solvent regions of most macromolecular crystals. The method is expected to be exact if all of the nonzero electron density lies within these expansion zones and the electron density outside of these expansion regions has a value that is uniformly zero. We have examined a few macromolecular crystals of known structure and have found that the experimental average coordinate of each asymmetric unit tends to lie within a few Å of those points in a unit cell that, when chosen as an origin, allow the largest spheres to be packed within the crystal lattice. (See also Hendrickson and Ward, 1976). Using these largest possible spheres, we have been able in one test case (nuclease from *Staphylococcus aureus*) to generate an accumulated diffraction pattern of a unit cell with enforced non-centrosymmetric crystal symmetry that has from 90-95% correlation with the amplitudes of the diffraction pattern calculated from the experimental coordinates. We are presently examining the general utility of this method for describing the contents of sparsely packed, non-centrosymmetric crystals and will report on these shortly.

We have described methods for the accurate conversion between a phased Fourier and spherical harmonic-Bessel representation. We have also shown that the resulting spherical harmonic-Besel representation may be applied to a relatively rapid automatic six-dimensional

overlap search that can utilize our previously described accurate target functions. While computation times for the exhaustive search appear to be substantially faster than previously exhaustive calculation schemes, and we have introduced improvements that result in accurate calculations at points on a 6-dimensional grid, the new problem that arises for a library-based search is one of rapid data storage and retrieval. Toward these ends, we are optimizing the file structures and the sorting schemes within our databases and we are carrying out test calculations for trial partial databases. We plain to convert more extensive molecular structural databases to lists of spherical harmonic coefficient for further tests. We also have briefly introduced an additional application of multi-center spherical harmonic-Bessel representations toward the description of the contents of an asymmetric unit of a sparsely packed, non-centrosymmetric crystal.

REFERENCES INCORPORATED BY REFERENCE

Arnold, C.M., Simon, S.I, and Friedman, J.M. *(to be submitted, Journal of Biological Chemistry).*

Buerger, M.J. *Vector Space*, Wiley & Sons, New York, 1959.

Chapman, M.S., Tsao, J., and Rossmann, M.G. (1992) *Acta Crystallographica*, **A48**, 301-312.

Cooley, J. and Tukey, J.W. (1965)*Mathematical Computation*, **19**, 297-301.

Crowther, R.A. (1972)*The Molecular Replacement Method*, M.G. Rossmann, Ed., Gordon & Breach, New York, pp. 173-178.

Dodson, E.J. (1985) *Molecular Replacement: Proceedings of the Daresbury Study Weekend, 15-16 February 1985*, P. A. Machin, Ed., SERC Daresbury Laboratory, Warrington, England, pp. 33-45.

Fitzgerald, P.M.D. (1988) *Journal of Applied Crystallography*, **21**, 273-278.

Friedman, J.M. (1997) *Protein Engineering*, **10**, 851-863.

Gradshteyn, I.S. and Ryzhik, I.M. (1980) *Table of Integrals, Series, and Products: Corrected and Enlarged Edition*, Academic Press, Orlando.

Harrison, R.W., Kourinov, I.V. and Andrews, L.C. (1994) *Protein Engineering*, **7**, 359-369.

Hendrickson, W.A. and Ward, K.B. (1976) *Acta Crystallographica* **A32**, 778-780.

Jones, T.A., Zou, J.-Y., Cowan, S.W. and Kjeldgaard, M. (1991) *Acta Crystallographica* **A47**, 110-119.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C. and Vakser, I.A. (1992) *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 2195-2199.

Kuntz, I.D., Meng, E.C. and Shoichet, B.K. (1994) *Accounts of Chemical Research*, **27**, 117-123.

Lattman, E.E. (1972) *Acta Crystallographica*, **B28**, 1065-1068.

Morse, P.M. and Feshbach, H. (1953) *Methods of Theoretical Physics*, p. 1467, McGraw-Hill, New York.

Navaza, J. (1987) *Acta Crystallographica*, **A43**, 645-653.

Navaza, J. (1990) *Acta Crystallographica*, **A46**, 619-620.

Nissink, J.W.M., Verdonk, M.L., Kroon, J., Mietzner, T., and Klebe, G. (1997) *Journal of Computational Chemistry*, **A32**, 638-645.

Podjarny, A.D. and Urzhumtsev, A. (1996) *Transactions of the American Crystallographic Association* **30**, 109-120.

Rossmann, M.G. ed. (1972) *The Molecular Replacement Method*, Gordon & Breach, New York.

Rossmann, M.G. (1990) *Acta Crystallographica*, **A46**, 73-82.

Ten Eyck, L.F. (1973) *Acta Crystallographica*, **A29**, 183-191.

Ten Eyck, L.F. (1977) *Acta Crystallographica*, **A33**, 486-492.

Tsao, J., Chapman, M.S., and Rossmann, M.G. (1992) *Acta Crystallographica*, **A48**, 293-301.

Thus, it can be appreciated that a computational method and an apparatus therefore have been presented which will facilitate the discovery of novel bio-active and/or therapeutic molecules, these methods rely on the use of a computational methods employing a general recursive method for determining the macromolecular crystallographic phases of molecules so as to recognize and predict ligand binding affinity.

Accordingly, it is to be understood that the embodiments of the invention herein providing for a more efficient mode of drug discovery and modification are merely illustrative of the application of the principles of the invention. It will be evident from the foregoing description that changes in the form, methods of use, and applications of the elements of the computational method and associated algorithms disclosed may be resorted to without departing from the spirit of the invention, or the scope of the appended claims.